ACQUITY GROUP
Part of Accenture Interactive

ADOBE EXPERIENCE MANAGER & EXTERNAL SEARCH PLATFORMS

Matthias Wermund, Senior Application Architect

WHY EXTERNAL SEARCH

Search is part of most implementation projects

- Most of today's web sites offer any type of search feature
- Search exists in various flavors and mixtures
  - **Site search**
  - **Typed search**
  - **Search as navigation**
  - **Relevance search**
  - **Location based search**
- AEM comes with its own search implementation
- "External search" in context of AEM means leveraging another platform, hosted outside of the AEM Author/Publish environments

WHY EXTERNAL SEARCH

Publish search vs. Author search

- Publish search
  - **End user accessible**
  - **Indexed content is in published state**
  - **High frequency access**
- Author search
  - **Internal AEM author search**
  - **Index must include unpublished content**
  - **Criteria can include additional content metadata**

- Both are fundamentally different use cases with different index lifecycle

WHY EXTERNAL SEARCH

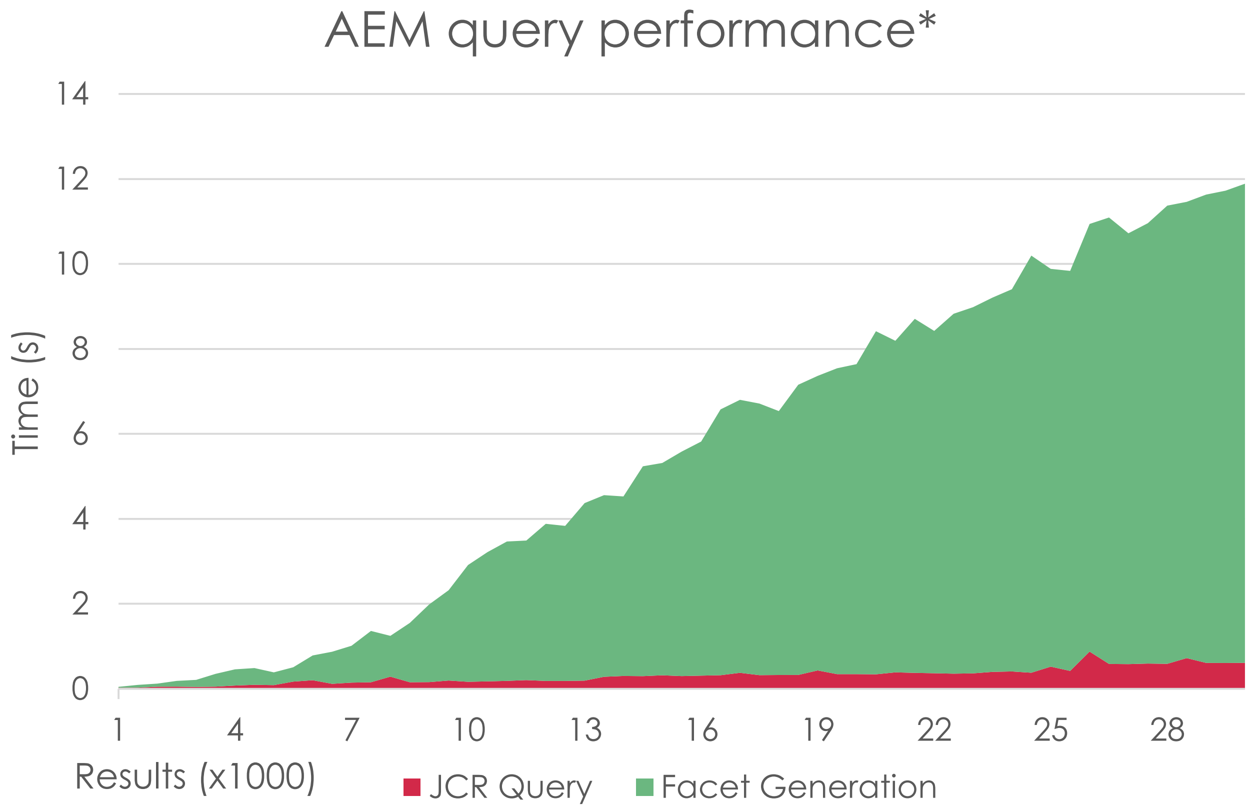
AEM standard search

- Part of AEM Java Content Repository (JCR) implementation
- AEM adds Predicate API layer
- Features
  - **Automatic index generation in all environments**
  - **Full-text, facetted search**
  - **Access restrictions based on repository access control lists (ACL)**
- Used behind the scenes for many AEM features

So, why not always use JCR search? Here are some reasons.

WHY EXTERNAL SEARCH

Performance issues of standard search with growing result size

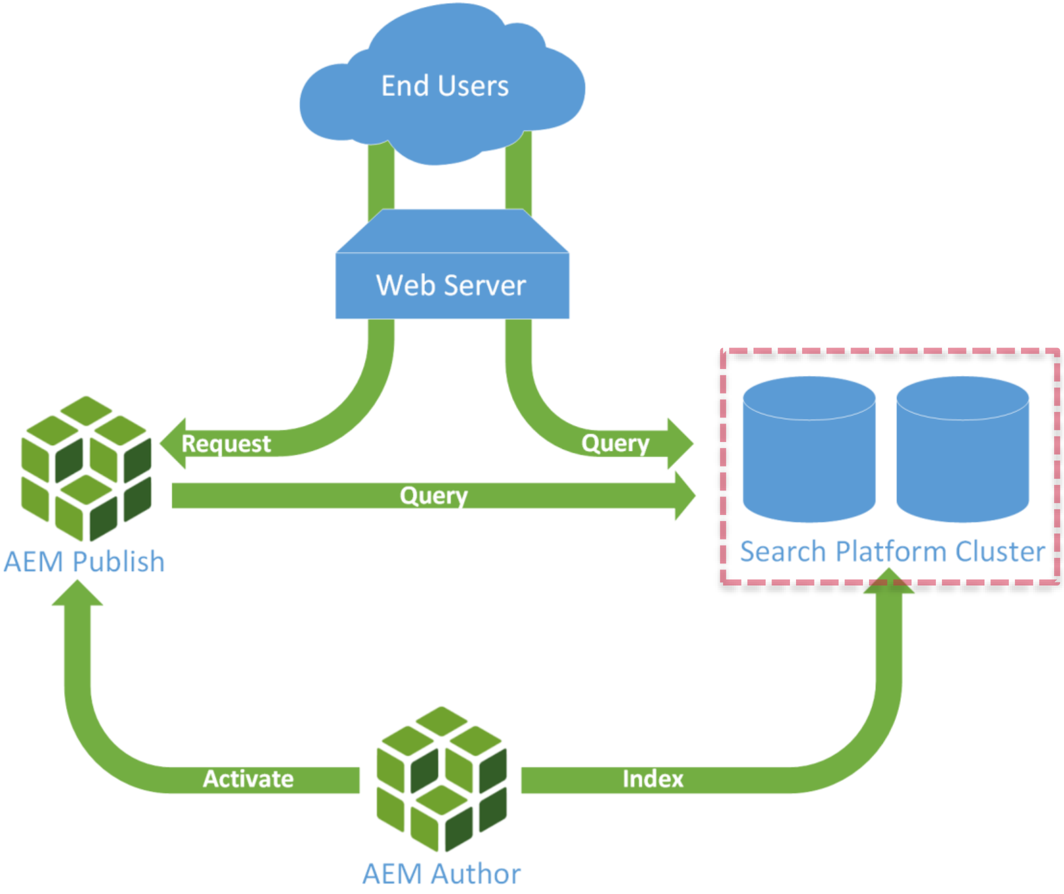


- Facet generation time increases linear with growing result size
- JCR query time not impacted the same, but increase is noticeable
- Search results are often impossible to cache due to high number of variations

* Synthetic content, full-text search, one facet, single requests

WHY EXTERNAL SEARCH

Scale independently of AEM



- External search platforms decouple the search infrastructure from AEM
- Search platform can scale independently from AEM, both horizontally and vertically
- Some platforms support cloud deployments, e.g. ZooKeeper for Apache Solr
- Client-side integration can fully eliminate query impact on AEM

WHY EXTERNAL SEARCH

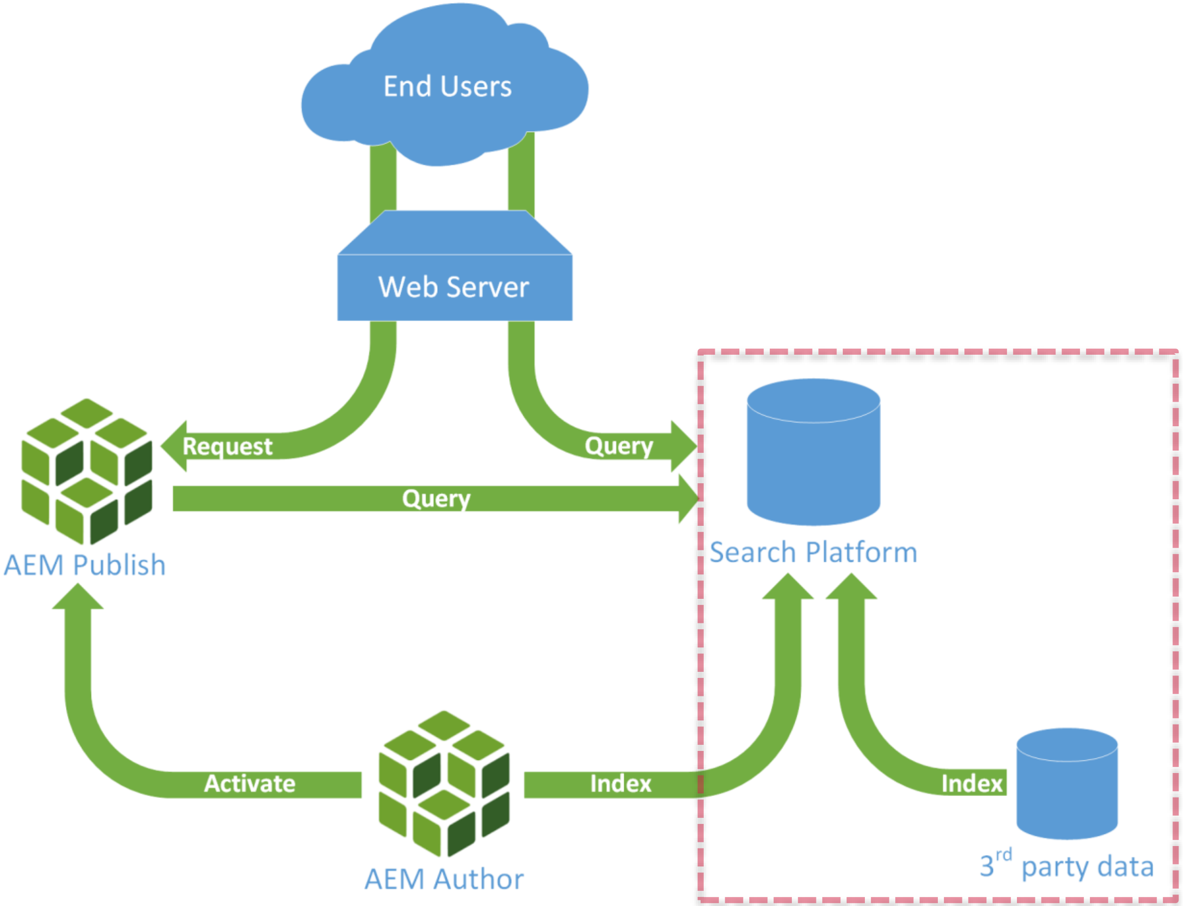
Extended feature offering

- External platforms provide functionality which AEM currently doesn’t bring out-of-box
- A few examples:
  - **Geospatial search**
  - **Dynamic relevance control**
  - **Index-based type-ahead**
  - **Index maintenance UI**

- More about various possible uses of external index data later.

WHY EXTERNAL SEARCH

Search multiple data sources at once



- Search can span multiple data sources besides AEM
- External platforms can join data from any number and type of source systems
- Users can query all data at once, with combined pagination, filters and relevance calculation

APACHE SOLR

Lucene-powered Open Source Search Platform

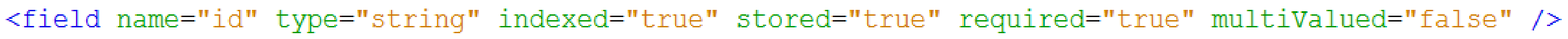
- Created initially by CNET, since 2006 open source
- Incorporates and extends Apache Lucene
  - **Supports distributed indexing and searching**
  - **Rich search capabilities**
  - **HTTP interface, JSON/XML/BIN formats**
  - **Integration clients**
  - **Standalone Java web application**
  - **Administration UI**

- Widely used on prominent sites
  wiki.apache.org/solr/PublicServers

APACHE SOLR

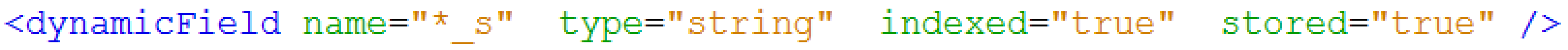
Index schema configuration

- Schema fields



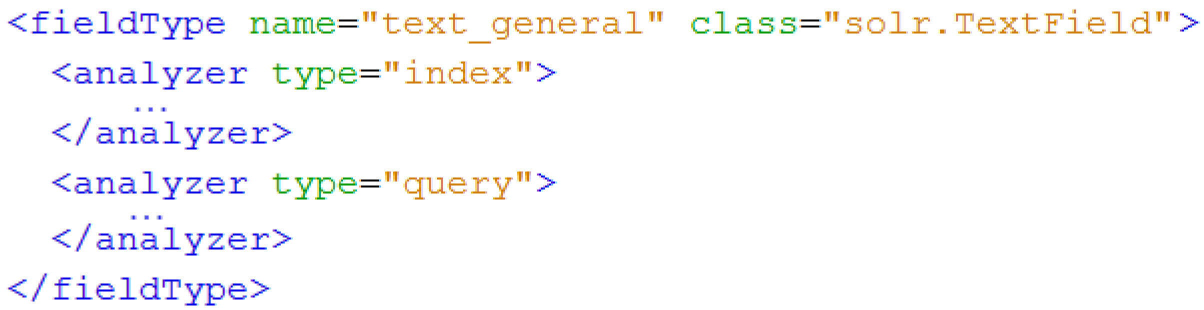
```
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false" />
```

- Dynamic fields



```
<dynamicField name="*_s"  type="string"  indexed="true"  stored="true" />
```

- Custom field types



```
<fieldType name="text_general" class="solr.TextField">
  <analyzer type="index">
     ...
  </analyzer>
  <analyzer type="query">
     ...
  </analyzer>
</fieldType>
```

INDEXING AEM CONTENT

Steps to create an external index

- The main challenge is **data extraction**
  - **When should the extraction process get initiated?**
  - **How to convert the AEM content tree structure into the index format?**
  - **And how to transfer the converted data to the external platform?**

- Once the external index is generated, **data querying** is a relatively easy step
  - **Query generation is highly specific to the use case**
  - **Most search platforms offer standard interfaces or Java libraries to integrate**

INDEXING AEM CONTENT

Integration patterns : Pull vs. Push

- Pull
  - **Content downloaded by external search platform**
  - **Platform needs trigger, e.g. scheduler**
  - **Data generation can use same rendering as for user requests**
- Push
  - **Data uploaded from AEM to external platform**
  - **Can happened immediately on modification**
  - **Requires to generate data standalone**

- Combination is possible – Example:
  - **On modification, AEM notifies search platform (Push)**
  - **Platform loads the modified content from AEM (Pull)**

INDEXING AEM CONTENT

Integration patterns (II) : Unstructured vs. Structured

- Unstructured
  - **No index-specific format**
  - **Metadata is extracted after loading**
  - **Least effort, end user rendering can be used**
- Structured
  - **Source data formatted to match search index structure**
  - **Leaner**
  - **Can carry different data than end user view**
  - **Requires structure generation in AEM**

- The typical **unstructured pull** data extraction is **crawling**

INDEXING AEM CONTENT

Introducing the EASE framework

- **EASE** = **E**xternal **A**EM **S**earch **E**xtension
- Primary goal of the framework is to reduce the complexity of integrating search platforms with AEM
- The indexing approach is **structured push** triggered by **content replication**
- Open source, available starting today
- For documentation, API, Maven dependencies & more see **github.com/mwmd/ease**

INDEXING AEM CONTENT

(Current) EASE framework features

- Supports generation of structured index data
- Binary asset indexing
- Integrates in AEM Author environment
- Incremental index updates triggered by AEM replication (push)
- Indexing of versioned content for scheduled replication
- Full index generation
- Generic integration with search platforms
- Apache Solr integration

INDEXING AEM CONTENT

EASE Maven modules

ease-core

- Provides indexing API
- Controls index modification
- Search platform agnostic
- OSGi bundle

ease-solr

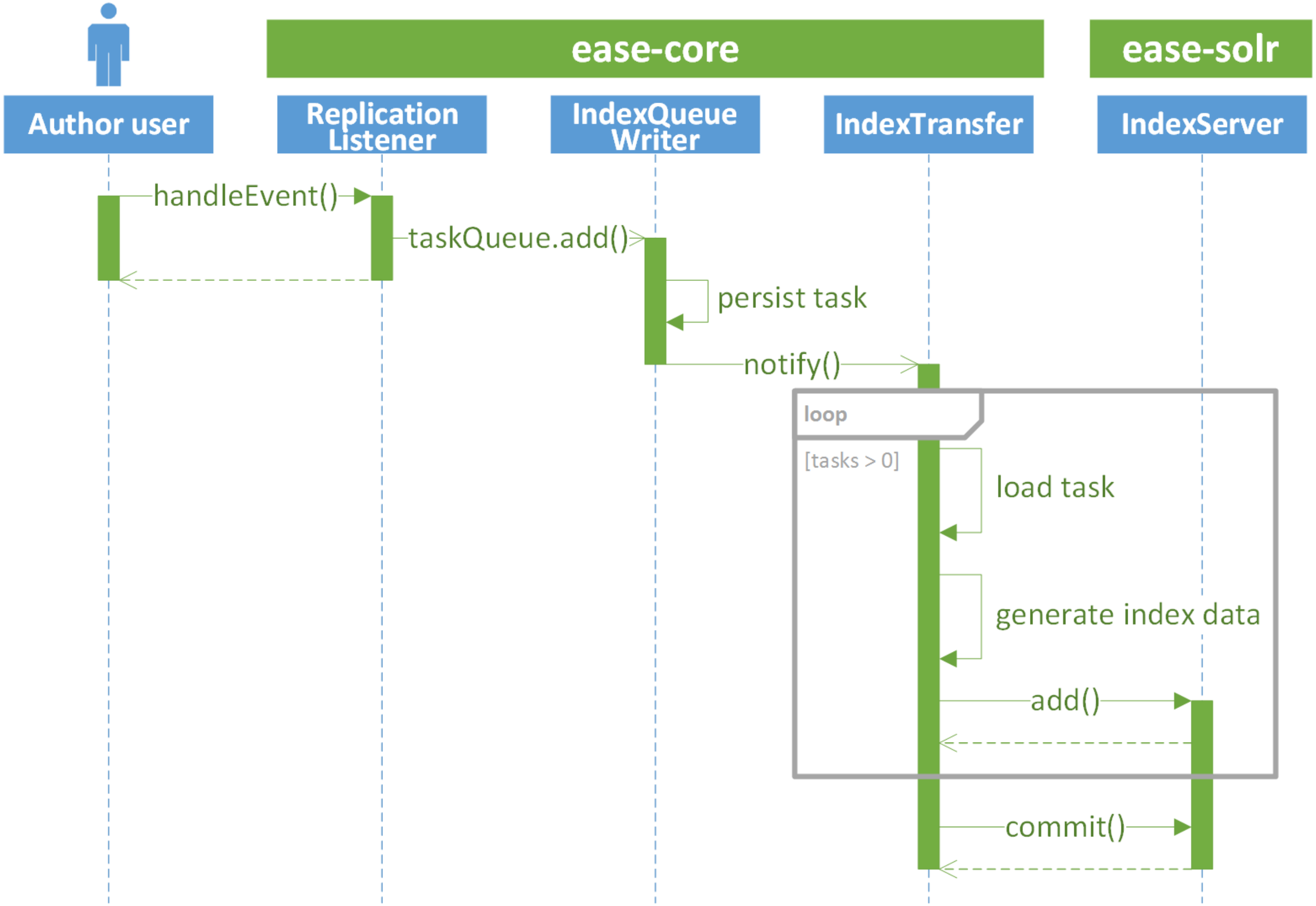
- Solr implementation core
- Provides SolrJ library to other consumer bundles
- OSGi bundle

ease-scr

- AnnotationProcessor for Felix Maven SCR Plugin
- Compile-time only dependency

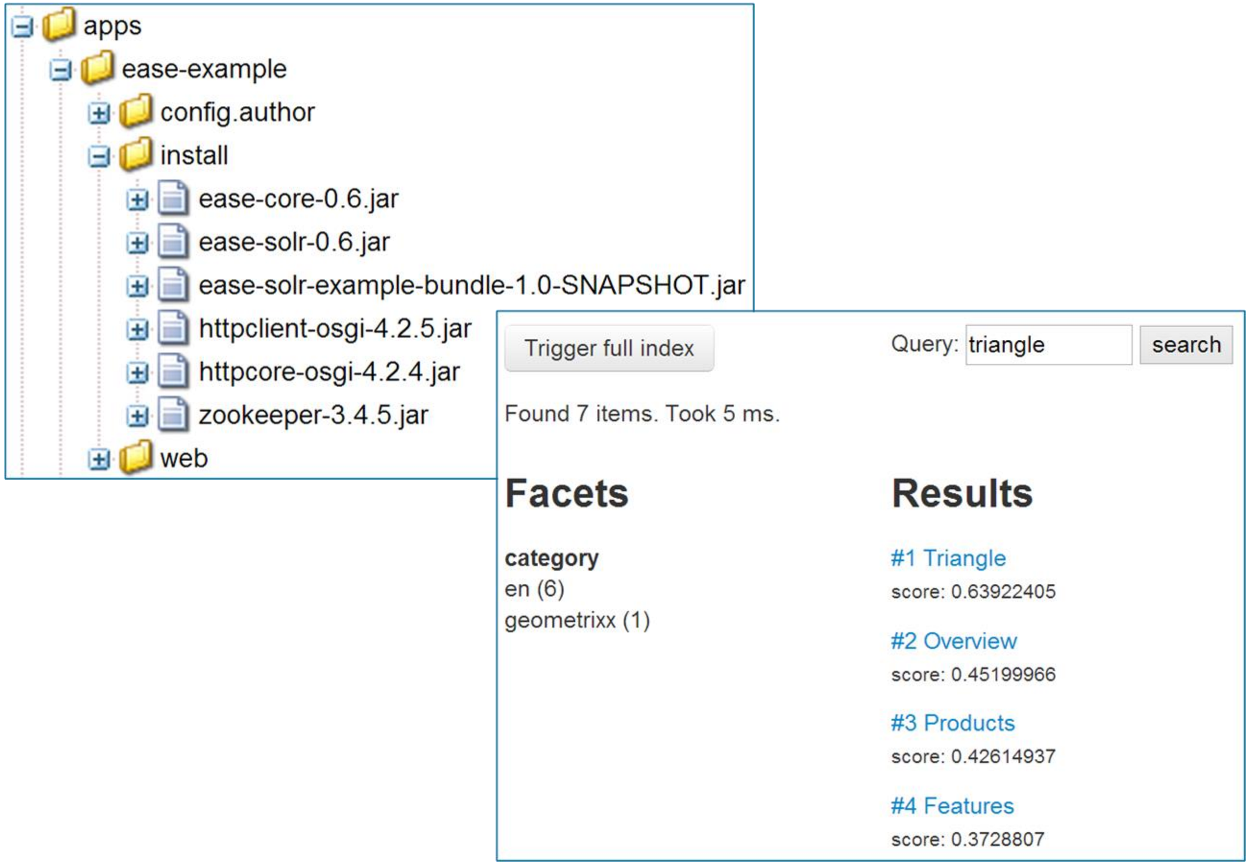
INDEXING AEM CONTENT

EASE index generation approach



INDEXING AEM CONTENT

Basic sample project: ease/example



```
apps
  ease-example
    config.author
    install
      ease-core-0.6.jar
      ease-solr-0.6.jar
      ease-solr-example-bundle-1.0-SNAPSHOT.jar
      httpclient-osgi-4.2.5.jar
      httpcore-osgi-4.2.4.jar
      zookeeper-3.4.5.jar
    web
```

Query: triangle

Found 7 items. Took 5 ms.

Facets

category
en (6)
geometrixx (1)

Results

#1 Triangle
score: 0.63922405

#2 Overview
score: 0.45199966

#3 Products
score: 0.42614937

#4 Features
score: 0.3728807

- Prerequisites
  - **AEM 5.6 Author**
  - **Apache Solr 4.4**
- Demonstrates use of EASE framework
- No configuration needed
- Uses Facets, full text search, relevance
- Available on GitHub

INDEXING AEM CONTENT

Steps to integrate EASE and Solr into project

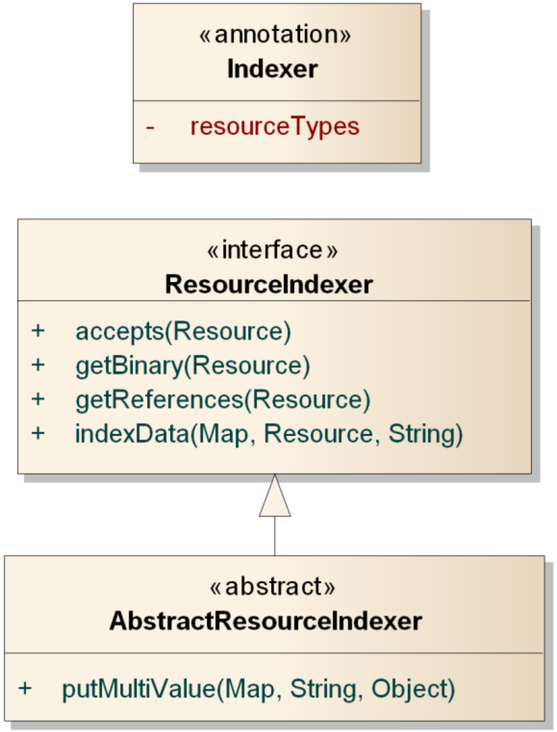
1. Include ease-core and ease-scr as Maven dependencies
2. Implement indexers matching your content
3. Create OSGi configurations:
   - **IndexService**
   - **SolrIndexServer**
4. Deploy to AEM Author:
   - **ease-core**
   - **ease-solr & dependencies**

When this is done, activated content will get automatically indexed.

INDEXING AEM CONTENT

Generation of index data with EASE

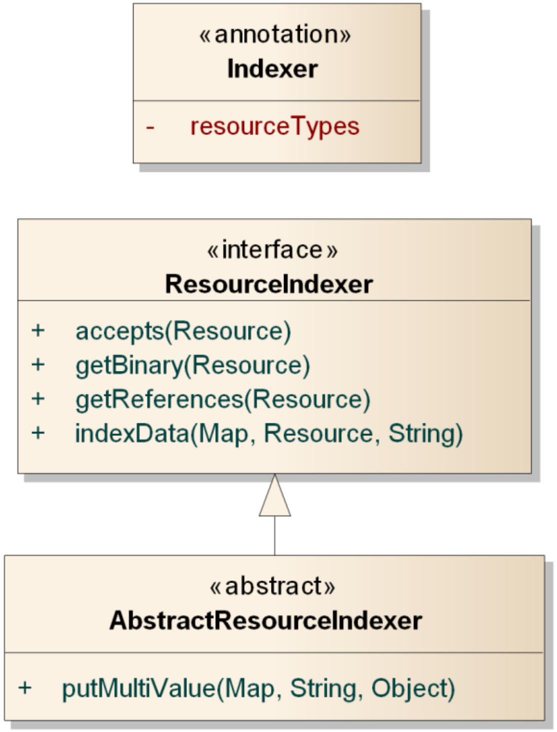
```
@Indexer(resourceTypes = "foundation/components/text")
public class TextIndexer extends AbstractResourceIndexer {

    @Override
    public void indexData(Map<String, Object> data,
            Resource res, String path) {

        ValueMap properties = res.adaptTo(ValueMap.class);
        String text = properties.get("text", String.class);
        putMultiValue(data, "text", text);
    }
}
```



INDEXING AEM CONTENT

Resolving content structure with EASE

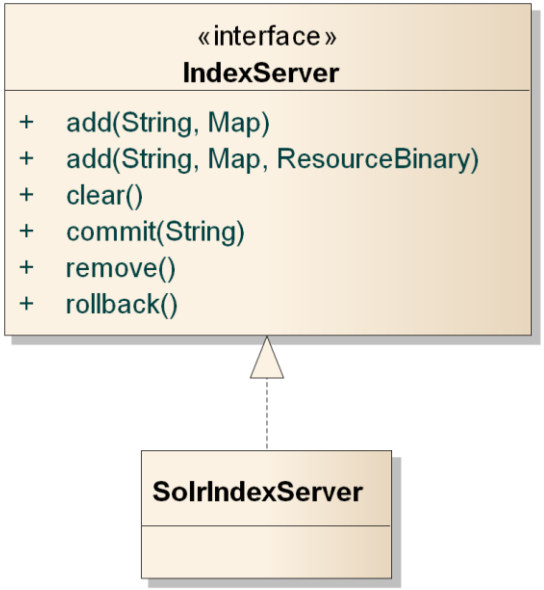
```java
@Indexer(resourceTypes = "geometrixx/components/contentpage")
public class PageIndexer extends AbstractResourceIndexer {

    @Override
    public List<ResourceReference> getReferences(Resource res) {

        Resource parRes = res.getChild("par");
        ResourceReference ref = new ResourceReference(parRes,
                "foundation/components/parsys");
        return Arrays.asList(ref);
    }
    // ...
}
```



INDEXING AEM CONTENT

Encapsulate handling of proprietary requests

- IndexServer handles all communication with search platform
- ease-core bundle doesn't provide platform specific implementation
- Implementation of IndexServer for Apache Solr in ease-solr bundle
- New connectors to additional platforms are only required to implement this interface



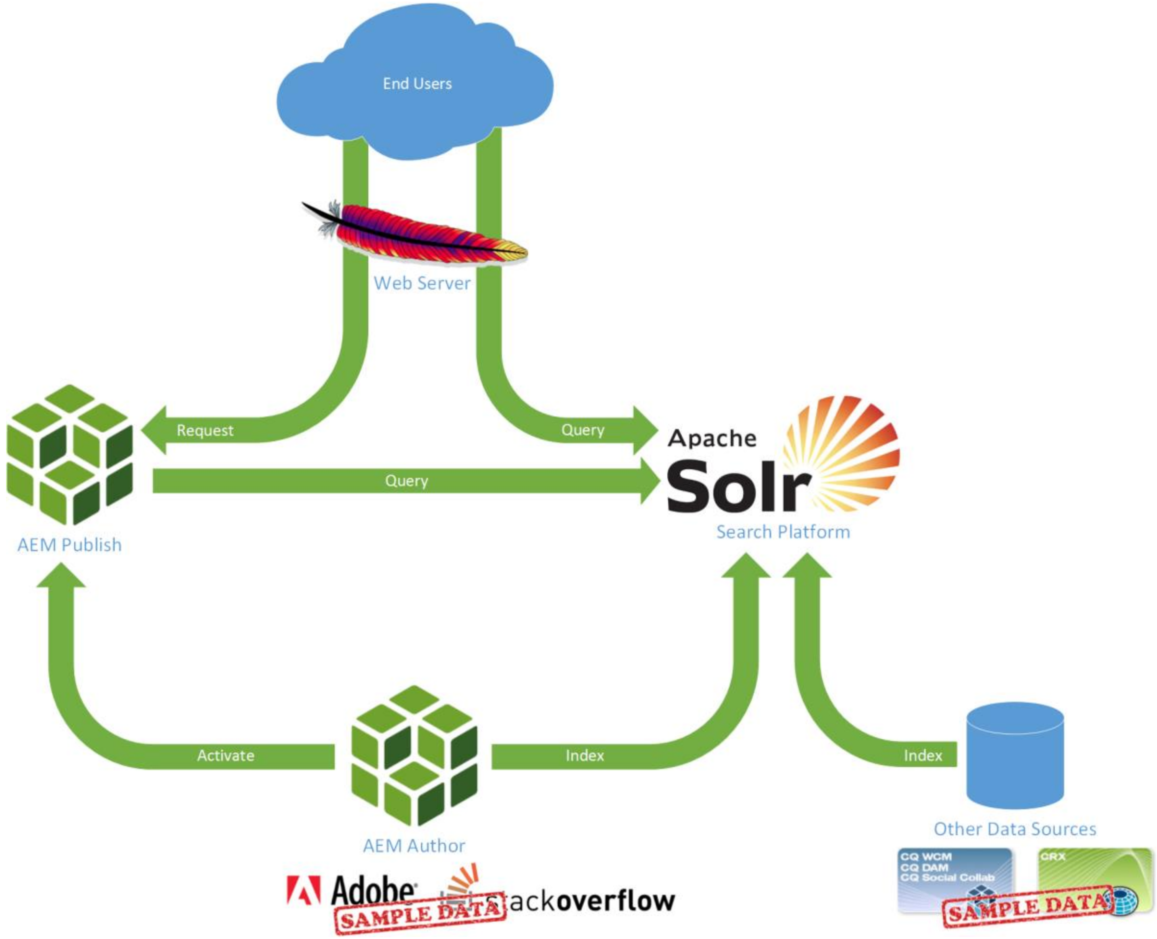
USING THE EXTERNAL INDEX

- When the data is indexed, it can get queried from custom components
- Leveraging platform specific features with proprietary clients
- While EASE currently focuses on simplifying the indexing, it helps with queries too
  - **EASE connector bundle per external search platform**
  - **Proprietary clients are provided by the bundle (SolrJ for Apache Solr)**

- In the following, an example implementation will walk through some use cases

USING THE EXTERNAL INDEX

Example implementation: AEM Know-How Database



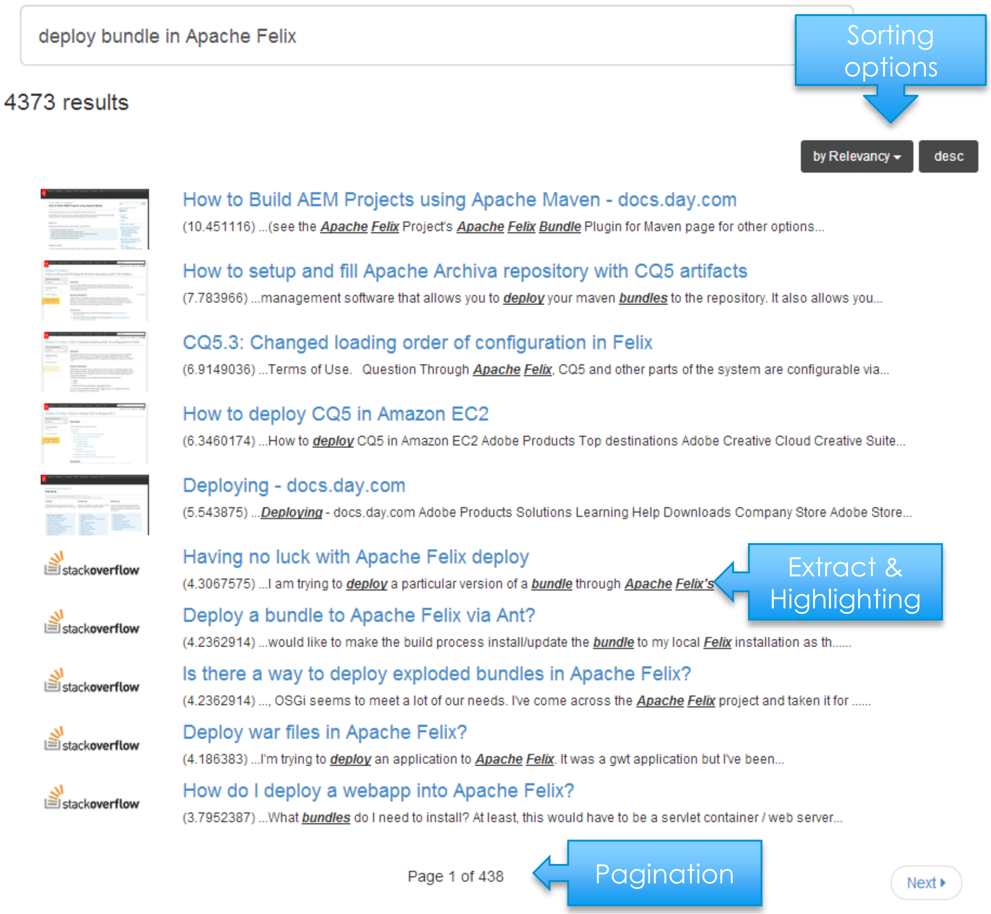
- Central search for AEM related information
- Uses EASE framework
- Server- and client-side queries
- 50,000 pages in AEM
  - **stackoverflow**
  - **Adobe offices**
- 3,000 external pages
  - **Adobe AEM doc**
  - **Adobe CRX doc**
  - **Marketing Cloud doc**

USING THE EXTERNAL INDEX

Example implementation: AEM Know-How Database

USING THE EXTERNAL INDEX

Full-text search



deploy bundle in Apache Felix

4373 results

How to Build AEM Projects using Apache Maven - docs.day.com
(10.451116) ...(see the Apache Felix Project's Apache Felix Bundle Plugin for Maven page for other options...

How to setup and fill Apache Archiva repository with CQ5 artifacts
(7.783966) ...management software that allows you to deploy your maven bundles to the repository. It also allows you...

CQ5.3: Changed loading order of configuration in Felix
(6.9149036) ...Terms of Use. Question Through Apache Felix, CQ5 and other parts of the system are configurable via...

How to deploy CQ5 in Amazon EC2
(6.3460174) ...How to deploy CQ5 in Amazon EC2 Adobe Products Top destinations Adobe Creative Cloud Creative Suite...

Deploying - docs.day.com
(5.543875) ...Deploying - docs.day.com Adobe Products Solutions Learning Help Downloads Company Store Adobe Store...

Having no luck with Apache Felix deploy
(4.3067575) ...I am trying to deploy a particular version of a bundle through Apache Felix's

Deploy a bundle to Apache Felix via Ant?
(4.2362914) ...would like to make the build process install/update the bundle to my local Felix installation as th......

Is there a way to deploy exploded bundles in Apache Felix?
(4.2362914) ..., OSGi seems to meet a lot of our needs. I've come across the Apache Felix project and taken it for ......

Deploy war files in Apache Felix?
(4.186383) ...I'm trying to deploy an application to Apache Felix. It was a gwt application but I've been...

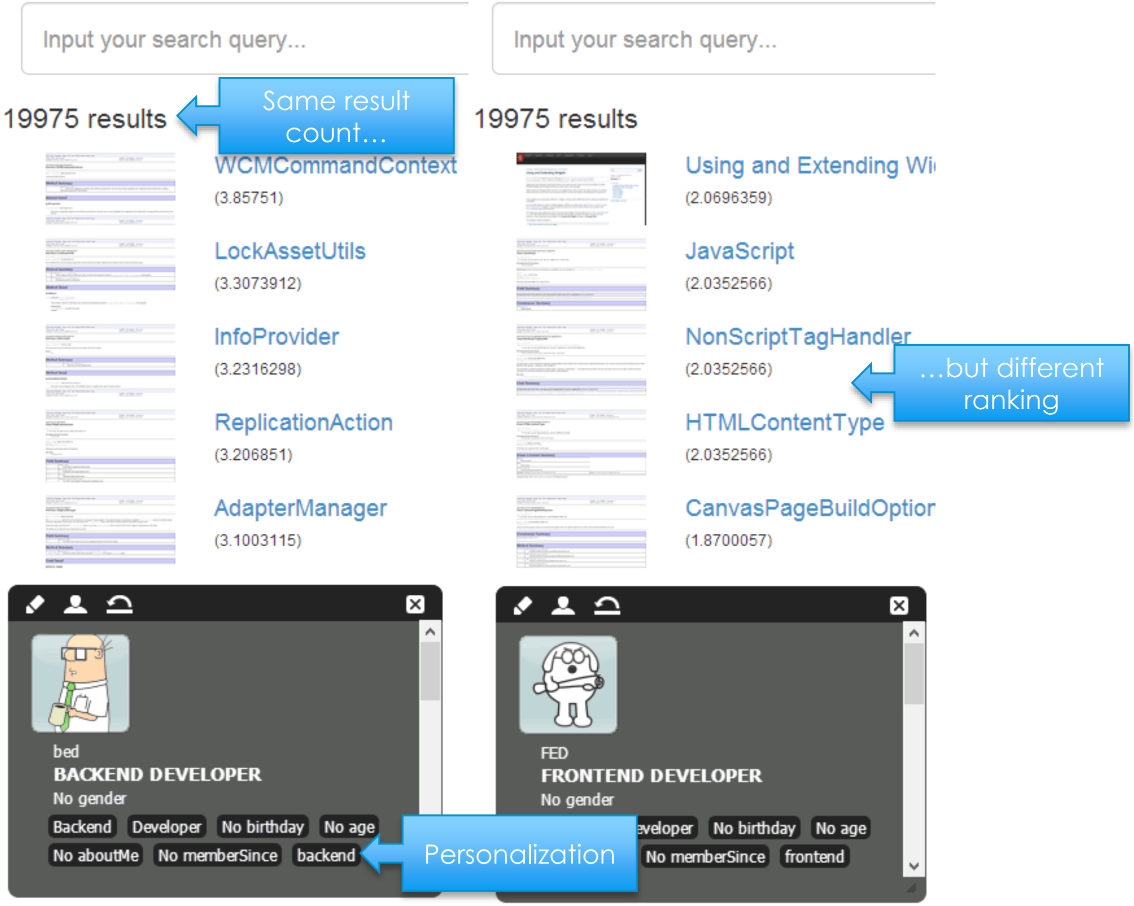
How do I deploy a webapp into Apache Felix?
(3.7952387) ...What bundles do I need to install? At least, this would have to be a servlet container / web server...

Page 1 of 438

- User-input text query
- Query-based ranking
- Generation of extracts and term highlighting
- Sorting on different fields

USING THE EXTERNAL INDEX

Boost manipulation / Personalization



Input your search query...

19975 results

Same result count…

WCMCommandContext
(3.85751)

LockAssetUtils
(3.3073912)

InfoProvider
(3.2316298)

ReplicationAction
(3.206851)

AdapterManager
(3.1003115)

Input your search query...

19975 results

Using and Extending Wi
(2.0696359)

JavaScript
(2.0352566)

NonScriptTagHandler
(2.0352566)

…but different ranking

HTMLContentType
(2.0352566)

CanvasPageBuildOptior
(1.8700057)

bed
BACKEND DEVELOPER
No gender
Backend | Developer | No birthday | No age
No aboutMe | No memberSince | backend

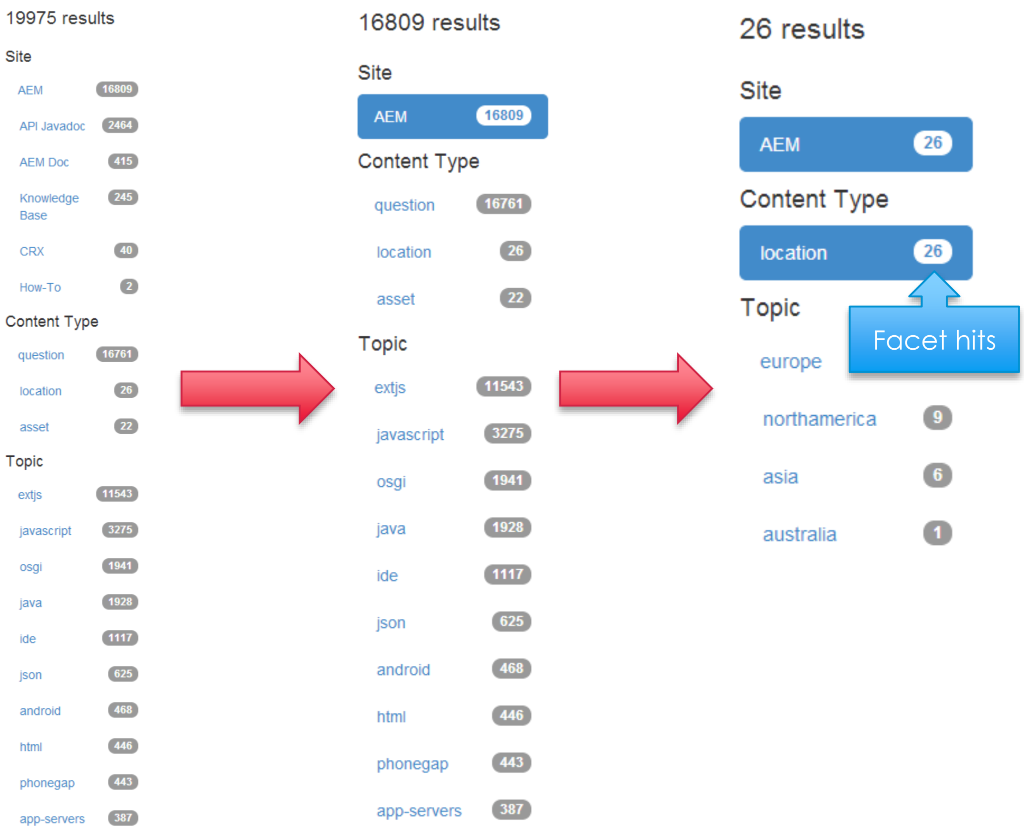
Personalization

FED
FRONTEND DEVELOPER
No gender
eveloper | No birthday | No age
No memberSince | frontend

- Manipulation of relevance calculation
- Boosting possible on
  - **Terms**
  - **Fields**
- Implementation can leverage user data to generate personalized result (Client Context)

USING THE EXTERNAL INDEX

Faceted search



19975 results

Site	
AEM	16809
API Javadoc	2464
AEM Doc	415
Knowledge Base	245
CRX	40
How-To	2

Content Type	
question	16761
location	26
asset	22

Topic	
extjs	11543
javascript	3275
osgi	1941
java	1928
ide	1117
json	625
android	468
html	446
phonegap	443
app-servers	387

16809 results

Site	
AEM	16809

Content Type	
question	16761
location	26
asset	22

Topic	
extjs	11543
javascript	3275
osgi	1941
java	1928
ide	1117
json	625
android	468
html	446
phonegap	443
app-servers	387

26 results

Site	
AEM	26

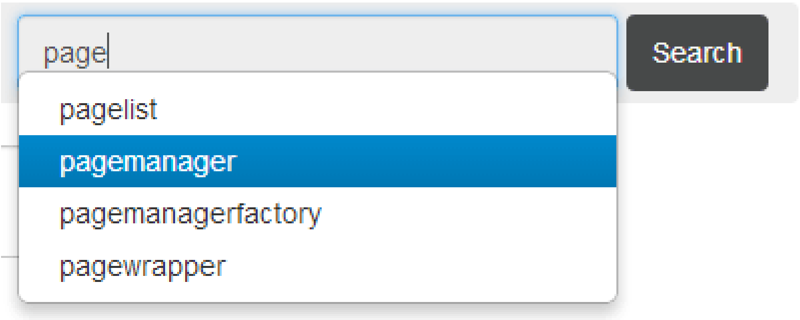
Content Type	
location	26

Topic	
europe	
northamerica	9
asia	6
australia	1

- Navigation via facets
- AND combination of multiple facet values
- Facet hit counts calculated based on current search result

USING THE EXTERNAL INDEX

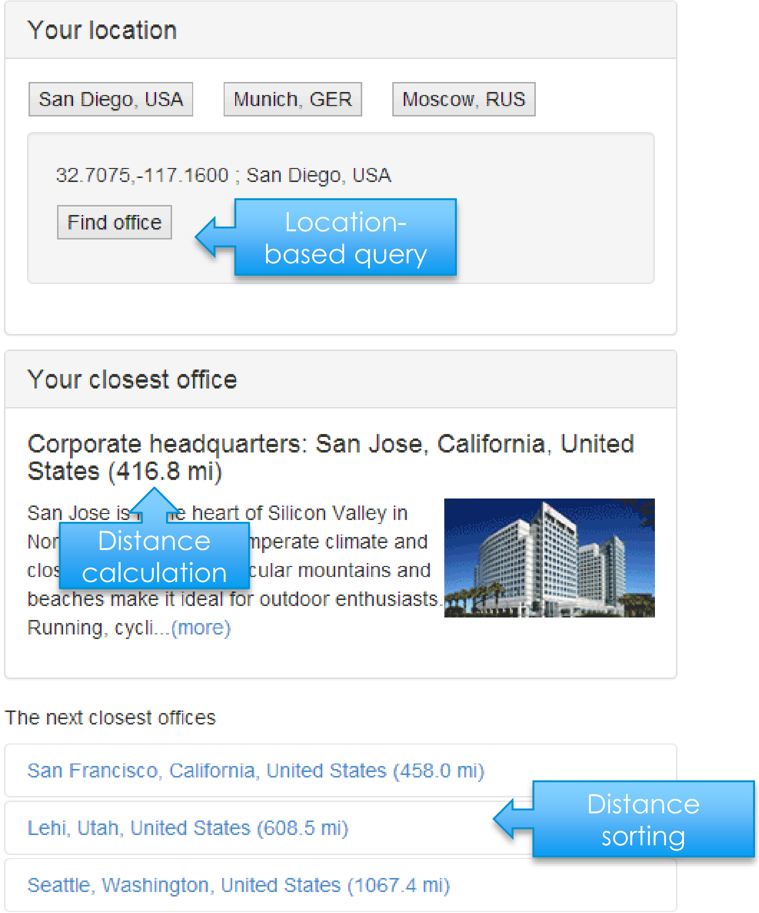
Type-ahead / Auto-complete / Auto-suggest



- Offer search term suggestions based on user input
- Highly configurable
  - **Index data**
  - **Dictionary**
  - **Query parsing**
- Client-side call to Apache Solr
- Can use a standard query or dedicated feature

USING THE EXTERNAL INDEX

Geospatial search



Your location

San Diego, USA | Munich, GER | Moscow, RUS

32.7075,-117.1600 ; San Diego, USA

Find office

Location-based query

Your closest office

Corporate headquarters: San Jose, California, United States (416.8 mi)

San Jose is the heart of Silicon Valley in Nor mperate climate and clos cular mountains and beaches make it ideal for outdoor enthusiasts. Running, cycli...(more)

Distance calculation

The next closest offices

San Francisco, California, United States (458.0 mi)

Lehi, Utah, United States (608.5 mi)

Seattle, Washington, United States (1067.4 mi)

Distance sorting

- Proximity search
- Distance calculation
- Sorting by distance
- Center of search is dynamic, can be based on user's location (Client Context)

REAL WORLD SEARCH IMPLEMENTATIONS

Handling aggregated page content

- Component rendering can include content maintained on other pages
- Aggregation logic could be easily mirrored in ResourceIndexer
- But: If the page-external content is modified, its **activation won't trigger re-indexing** of the aggregating page

- Use case:
  - **Inherited paragraph system**
  - **Reference component**

- Mitigation options:
  - **Content strategy: Index only standalone, unique page content**
  - **Use WCM ReferenceSearch to find and re-index references**
    - **Dangerous: reference loops, cascading re-indexing**

REAL WORLD SEARCH IMPLEMENTATIONS

Permission Sensitive Search

- External platforms have no information about AEM roles and permissions
- All index items are visible to everyone by default
- In some use cases, **access to parts of the index must be restricted**

- Use case:
  - **Closed User Groups**

- Mitigation options:
  - **Check access rights for all items on the current result page at runtime**
    - **Will break pagination information, type-ahead suggestions**
    - **Performance hit**
  - **Export effective role permissions as part of index item metadata**
    - **Add filter for current user's role to all queries or into search platform**
    - **Requires re-indexing of content on ACL changes**
    - **Only practical with a limited number of roles**

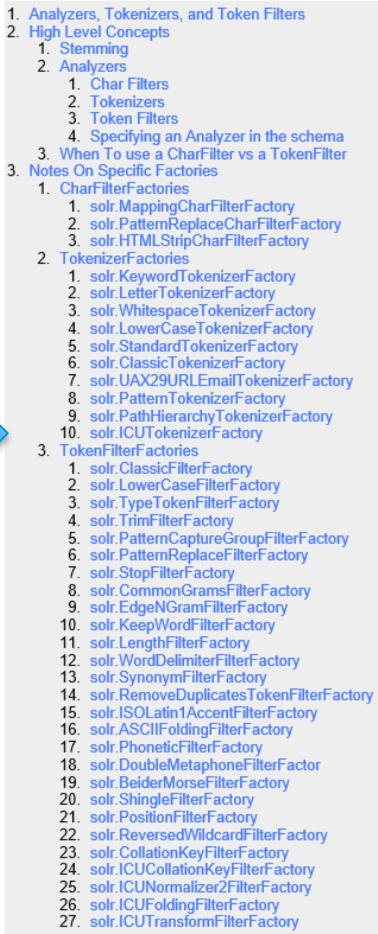
REAL WORLD SEARCH IMPLEMENTATIONS

Index tuning

- Interpretation of raw index data dependent on search technology and configuration
- Powerful platforms offer deep level of index configuration
- Tuning of search behavior means significant effort

- Use case:
  - **Full text query and content parsing**
  - **Type-ahead suggestions**
  - **Result relevance calculation**

List of available text processors for Solr



1. Analyzers, Tokenizers, and Token Filters
2. High Level Concepts
   1. Stemming
   2. Analyzers
      1. Char Filters
      2. Tokenizers
      3. Token Filters
      4. Specifying an Analyzer in the schema
   3. When To use a CharFilter vs a TokenFilter
3. Notes On Specific Factories
   1. CharFilterFactories
      1. solr.MappingCharFilterFactory
      2. solr.PatternReplaceCharFilterFactory
      3. solr.HTMLStripCharFilterFactory
   2. TokenizerFactories
      1. solr.KeywordTokenizerFactory
      2. solr.LetterTokenizerFactory
      3. solr.WhitespaceTokenizerFactory
      4. solr.LowerCaseTokenizerFactory
      5. solr.StandardTokenizerFactory
      6. solr.ClassicTokenizerFactory
      7. solr.UAX29URLEmailTokenizerFactory
      8. solr.PatternTokenizerFactory
      9. solr.PathHierarchyTokenizerFactory
      10. solr.ICUTokenizerFactory
   3. TokenFilterFactories
      1. solr.ClassicFilterFactory
      2. solr.LowerCaseFilterFactory
      3. solr.TypeTokenFilterFactory
      4. solr.TrimFilterFactory
      5. solr.PatternCaptureGroupFilterFactory
      6. solr.PatternReplaceFilterFactory
      7. solr.StopFilterFactory
      8. solr.CommonGramsFilterFactory
      9. solr.EdgeNGramFilterFactory
      10. solr.KeepWordFilterFactory
      11. solr.LengthFilterFactory
      12. solr.WordDelimiterFilterFactory
      13. solr.SynonymFilterFactory
      14. solr.RemoveDuplicatesTokenFilterFactory
      15. solr.ISOLatin1AccentFilterFactory
      16. solr.ASCIIFoldingFilterFactory
      17. solr.PhoneticFilterFactory
      18. solr.DoubleMetaphoneFilterFactor
      19. solr.BeiderMorseFilterFactory
      20. solr.ShingleFilterFactory
      21. solr.PositionFilterFactory
      22. solr.ReversedWildcardFilterFactory
      23. solr.CollationKeyFilterFactory
      24. solr.ICUCollationKeyFilterFactory
      25. solr.ICUNormalizer2FilterFactory
      26. solr.ICUFoldingFilterFactory
      27. solr.ICUTransformFilterFactory

ADOBE EXPERIENCE MANAGER & EXTERNAL SEARCH PLATFORMS

Questions?

matthias.wermund@acquitygroup.com

Thank you!